

08/21/00  
JC860 U.S. PTO

08-22-00

A

JC864 U.S. PTO  
09/642575

08/21/00

# UTILITY PATENT APPLICATION TRANSMITTAL

Attorney Docket No.

P00-3223

First Inventor

Richard D. Gunlock

Title

Apparatus and Method for  
Dynamic Command Queue  
Depth Adjustment for Storage  
Area Network Nodes

Express Mail Label No.

EK029467148US

## APPLICATION ELEMENTS

Assistant Commissioner for Patents  
Box Patent Application  
Washington, DC 20231

1. ☒ Fee Transmittal Form 5.
2. ☒ Specification [ total pages 38 ]
  - Descriptive title of the Invention
  - Cross References to Related Applications
  - Statement Regarding Fed sponsored R&D
  - Reference to Microfiche Appendix
  - Background of the Invention
  - Brief Summary of the Invention
  - Brief Description of the Drawings
  - Detailed Description
  - Claim(s)
  - Abstract of the Disclosure
3. ☒ Drawings(g) [ total sheets 12 ]
4. ☒ Oath or Declaration [signed] [ total sheets 1]
  - a. ☒ Newly executed (original)
  - b. ☐ Copy from prior appl. (37 C.F.R. § 1.63(d))
  - i. ☐ **DELETION OF INVENTOR(S)**  
Signed statement attached deleting inventor(s) named in prior application, see 37 C.F.R. §§ 1.63(d)(2) and 1.33(b).

- ☐ Microfiche Computer Program (*Appendix*)
- 6. Nucleotide/Amino Acid Sequence (if applicable)
  - a. ☐ Computer Readable Copy
  - b. ☐ Paper Copy (identical to computer copy)
  - c. ☐ Statement verifying identity of said copies

## ACCOMPANYING APPLICATION PARTS

7. ☒ Assignment Papers (coversheet/document(s))
8. ☐ 37 C.F.R. § 3.73(b) Statement (when there is an assignee)
  - ☒ Power of Attorney
  - ☐ Copies of
9. ☐ English Translation
10. ☐ IDS & Form 1449
11. ☐ Preliminary Amendment
12. ☒ Return Receipt Postcard (MPEP 503)
13. ☐ Small Entity ☐ Statement filed in prior application—Status still proper and desired
14. ☐ Certified Copy of Priority Document(s)
15. ☒ Other: Certificate of Mailing by Express Mail

NOTE FOR ITEMS 1 & 13: IN ORDER TO BE ENTITLED TO PAY SMALL ENTITY FEES, A SMALL ENTITY STATEMENT IS REQUIRED (37 C.F.R. § 1.27), EXCEPT IF ONE FILED IN A PRIOR APPLICATION IS RELIED UPON (37 C.F.R. § 1.28)

## 16. If a CONTINUING APPLICATION,

- ☐ Continuation ☐ Divisional ☐ Continuation-in-part (CIP) of prior application No.:   /
- Prior application information: Examiner:    Group/Art Unit:

FOR CONTINUATION OR DIVISIONAL APPS only: The entire disclosure of the prior application, from which an oath or declaration is supplied under Box 4b, is considered a part of the disclosure of the accompanying continuation or divisional application and is hereby incorporated by reference. The incorporation can only be relied upon when a portion has been inadvertently omitted from the submitted application parts.

## 17. CORRESPONDENCE ADDRESS

- ☒ Customer Number of Bar Code Label 25235 or ☒ Correspondence address below

Name	William J. KUBIDA, Esq.				
	Hogan & Hartson, LLP				
Address	1200 17 <sup>th</sup> Street				
	Suite 1500				
City	Denver	State	CO	ZIP	80202
Country	US	Telephone	(719) 448-5900	Fax	(719) 448-5922

Name (Print/Type)	Steven K. Barton	Registration No.	36,445
(Signature)	<i>Steven K. Barton</i>	Date	Aug 17, 2000

## IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:

Richard D. Gunlock

Serial No. \_\_\_\_\_

Filed: Herewith

For: Apparatus and Method for Dynamic  
Command Queue Depth Adjustment for  
Storage Area Network Nodes

Group Art Unit: \_\_\_\_\_

Examiner: \_\_\_\_\_

CERTIFICATE OF MAILING BY EXPRESS MAIL

## BOX PATENT APPLICATION

Assistant Commissioner of Patents and Trademarks  
Washington, D.C. 20231

Sir:

The undersigned hereby certifies that the following documents:

1. Utility Patent Application Transmittal;
2. Fee Transmittal and \$ 690 filing fee;
3. Utility Patent Application;
4. Executed Declaration;
5. Executed Power of Attorney by Assignee
6. 12 sheets of drawings;
7. Recordation Form Cover Sheet PTO 1595 with Executed Assignment and Recording Fee of \$40.00;
8. Return postcard; and
9. Certificate of Mailing By Express Mail

relating to the above application, were deposited as "Express Mail", Mailing Label No. EK029467148US with the United States Postal Service, addressed to The Commissioner of Patents and Trademarks, Washington, D.C., 20231, August 21, 2000.

August 21, 2000  
Date

Sheri Olsen  
Mailer

Aug 21, 2000  
Date

Steven K. Barton  
Steven K. Barton, Reg. No. 36,445  
Hogan & Hartson LLP  
One Tabor Center  
1200 17<sup>th</sup> Street, Suite 1500  
Denver, Colorado 80202  
Tel (719) 448-5909  
Fax (303) 899-7333

001280 54524960

**APPARATUS AND METHOD FOR DYNAMIC COMMAND QUEUE DEPTH  
ADJUSTMENT FOR STORAGE AREA NETWORK NODES**

**RELATED APPLICATIONS**

5 This application is related to a simultaneously  
filed application entitled " Data Structure and Method  
for Tracking Network Topology in a Fibre Channel Port  
Driver ", identified as Serial No. [XX/XXX,XXX] and  
attorney docket number P00-3221. This application is  
also related to a simultaneously filed application  
entitled "Method for Recovery of Paths Between Storage  
Area Network Nodes with Probationary Period and  
10 Desperation Repair", identified as Serial No.  
[XX/XXX,XXX] and attorney docket number P00-3222.

**FIELD OF THE INVENTION**

15 The invention relates to the field of computer  
networking. In particular, the invention relates to  
the field of nodes and switches for Fibre Channel  
storage-area networks. With more particularity, the  
invention relates to data structures for use by nodes  
in dynamically tracking network resources and  
configuration so as to properly encode and route  
20 packets over the network.

**NATURE OF THE PROBLEM**

25 Most modern computer networks, including switched  
and arbitrated-loop fibre-channel networks, are packet  
oriented. In these networks, data transmitted between  
machines is divided into chunks of size no greater  
than a predetermined maximum. Each chunk is typically  
packaged with a header and a trailer into a packet for

transmission. In Fibre-Channel networks, packets are known as Frames.

A Fibre-Channel network having at least one switch is a switched Fibre-Channel fabric. A Fibre-Channel switch is a routing device generally capable of receiving frames, storing them, decoding destination information from headers, and forwarding them to their destination or another switch further along a path toward their destination. A network interface for connection of a machine to a Fibre Channel fabric is known as an N\_port, and a machine attached to a Fibre-Channel network is known as a node. Nodes may be computers, or may be storage devices such as RAID systems. An NL\_port is an N\_port that supports additional arbitration required so that it may be connected either to a Fibre Channel Fabric or to a Fibre Channel Arbitrated Loop, and an L\_port is a network interface for connecting a node to a Fibre Channel Arbitrated Loop.

A device including an N\_port, L\_port, or an NL\_port together with hardware for high-speed connection to a machine is a fibre channel host bus adapter (physical HBA). For example, a physical HBA may comprise a printed circuit card having one or more NL\_ports communicating through a PCI bus interface to an edge connector for connection to a PCI bus of a machine. A physical HBA may, but need not, also incorporate a processor for controlling its ports and its connection to the machine.

A Fibre Channel Switched Fabric may connect one or more Fibre Channel Arbitrated Loops.

In a switched fibre channel fabric, there may be more than one possible path, or sequence of links, loops, switches, routers, etc. that may be traversed by a frame, between two nodes. Multiple paths may be intentional, providing extra capacity or redundancy to protect against switch, node, or line failures, or may be unintentional consequences of network topology.

Multiple paths between two nodes may also be provided through multiple fibre channel arbitrated loops. For example, an initiator node may have two NL\_ports, one connected to each of two fibre channel arbitrated loops. If each of these loops connects to an NL\_port of a target node, then multiple paths from the initiator to the target node exist. This could provide redundancy should a failure occur on one of the arbitrated loops.

Fibre Channel storage-area-network (SAN) nodes and switches, especially network nodes having multiple ports, must keep track of a variety of information about the network and resources available over the network. This information is used by each node to format and properly route frames onto and over the network.

Typically, a program references files in storage by passing a command block to an operating system. At this level, the command block may reference files by name and device name, or by handle for files already opened. The operating system passes block I/O requests derived from the command block to an appropriate driver. The command block includes a command field, a file name field, and a drive name or number field in addition to other fields. The block

5           The device driver typically converts the block  
I/O requests into a sequence of one or more device  
level commands. Once the driver attempts to execute  
each command, it returns to the operating system a  
response having status information, and other  
10 information possibly including data read from storage.  
Information from the response may then be transferred  
to the program.

Typically, the operating system passes block I/O requests to the device driver with a device tag identifying the specific device intended to perform the desired operation. This tag may comprise a referenced device name, handle, or SCSI nexus, where a SCSI nexus includes bus identity, target device number, and logical unit number.

In particular, the driver must translate device tags, into a multilevel address as required to reach the indicated device. In a Fibre Channel context, that multilevel address field must include valid

Destination Identification (D\_ID) fields and routing information field for each frame. A command frame may specify a specific logical unit. There may also be additional destination address fields such as those in association headers; association headers permit addressing multiple devices or processes through a single fibre channel port.

The driver, especially a multiport driver servicing multiple ports, must therefore determine an appropriate destination and routing for each frame required to implement a command, and transmit those frame over a port appropriate for that routing. It is desirable that these translations and assignments be done quickly and accurately using network topology information maintained in a local topology database.

The command and data frames transmitted by a node in implementing an I/O command, together with any frames transmitted by another node in response to those command and data frames, is known as an exchange.

As nodes, switches, and links are added to or removed from the network, any local topology database must be updated to reflect valid devices on the network, and valid paths through the network to those devices. Nodes also may determine one or more paths of the valid paths to a given device to be an "active" path. An active path is a path that may be used for exchanges.

The Fibre Channel specifications define Class 1 and Class 4 services to be virtual-circuit, or connection, based services between pairs of nodes.

Packets of a given sequence in these services must arrive in-order with respect to other packets of the same sequence. The specifications presume that frames transiting between nodes of each pair follow a virtual circuit between the nodes -- all following the same path through the network from node to node and arriving in-order.

Links, loops, and switches of a network may fail. Fibre channel networks may provide more than one path between a pair of nodes. Multiple, redundant, paths provide redundancy to allow continued communications between a pair of nodes should one or sometimes more, but not all, links, loops, or switches, and therefore paths through the network, fail.

Some existing fibre-channel systems can recognize failure of a path, switching traffic between a pair of nodes to an alternate path if one exists. This is known as failover of a path. Unfortunately, switching Class 1 and Class 4 fibre channel connections, and associated network traffic, to alternate paths is known to cause considerable delay to ongoing exchanges -- on the order of seconds. These delays are caused in part by the need to ensure that Class 1 and Class 4 frames arrive at their destination in-order; time is allowed for flushing in-transit frames before frames are transmitted on the alternate path.

Further delay may occur if redundant controllers having access to the target device must transfer ownership of a logical unit as part of a path failover.



While a link or switch of a fibre channel network may fail, it may also be repaired. It is desirable that paths associated with repaired links, loops, or switches be dynamically returned to use to restore network redundancy and capacity.

Loops, links or switches may develop intermittent problems, where they may alternate between repaired and failed conditions. Considerable delay of network traffic can be encountered if paths through an intermittent device are used, since the delays associated with switching traffic to repaired paths, thence back to alternate paths upon the next failure may be cumulative. Paths may also be reported as repaired but actually have higher error rates. High error rates cause delay as frames are retransmitted as errors are detected; high error rates may also cause a path to be considered to have failed. It is therefore desirable to control use of repaired paths such that extended delay does not occur.

It is known that many nodes, including RAID storage subsystems, have the ability to queue multiple commands in a command queue for subsequent execution. For example, a RAID system may queue several read or write commands, received from one or more machines. Once queued, these commands are executed from the queue to or from cache, or to or from disk, in an order depending on availability of data in cache, disk availability and disk rotation. With proper interlocks, execution may often be in an order different from that in which the commands were received. Commands placed in a command queue of a node by an initiator node and not yet completed are pending commands. For this purpose, commands that are

aborted, including those aborted for timeouts, are considered completed.

Commands that may be queued in these devices may include commands from multiple processes, or threads, running on a single node having one or more processors. For example, a transaction-processing system may have several processes running, each process requiring access to a different record of a database on a RAID system, all requesting access to the database at about the same time. Each process may create read, write, lock, or unlock commands for the database. Commands may also be queued from multiple nodes, where each node is running one or more processes or threads that require access to the device. Queuing and execution of each of these commands requires an exchange of frames between the machine and the device.

The maximum number, or queue depth, of commands that may be pending in any one device at any one time is finite and characteristic of the device. Each node may, but need not, know the queue depth of the device nodes it is attempting to use, and may throttle its commands appropriately. Typically, each node originating commands and transmitting them to the command queue of a device is unaware of the commands originated by other nodes and assumes it has the full queue depth of the device available to it.

If a node assumes that the full queue depth of a device is available to it, the queue depth of the device may be exceeded by commands from multiple nodes. When this happens, the excess commands are refused by the device and must be retransmitted later;

the need for delay and retransmission degrades overall system performance. Conversely, if a node assumes that a device has only a small queue depth available to it, commands may be throttled unduly, such that overall system performance is impaired. It is desirable to adjust the maximum number of commands from each node transmitted to a device command queue in light of the command traffic to that device from other nodes of the network.

Manually tuned allocations of queue depth available to individual initiator nodes of a storage area network is possible, but is laborious and error prone. Further, it is known that the load on a system may vary from time to time throughout a day; it is desirable to change queue depth allocations as load changes for optimum performance.

### SOLUTION TO THE PROBLEM

A driver for one or more Fibre Channel N\_ports, L\_ports, or NL\_ports maintains a local network topology database. This database includes linked lists of host bus adapter (HBA) port records, target node records, and device records.

HBA port records correspond to individual physical N\_ports, L\_ports or NL\_ports and are linked to node records through linked lists of node links. Similarly, node records are linked to device records through linked lists of device links, and devices are linked to node links through linked lists of path links. Additional HBA port records may be added as ports are discovered or as hot-pluggable ports are added to the system.

Each node record has information pertinent to a particular port of a target node of the fibre channel network. A target node may have more than one port, and may therefore have more than one node record.

5 Similarly, each device record has information  
pertinent to a particular device that can be accessed  
from one or more nodes. Each path link has  
information, including status information, pertinent  
to a particular path through a fibre channel fabric or  
10 arbitrated loop between a physical HBA port and a  
physical node or device of the network.

In order to reduce command queue overflow errors,  
a command queue depth variable is maintained for each  
node record. Initially, this is set to match a  
15 maximum queue depth determined from the type of node  
represented by the record. This queue depth variable  
is adjusted downwardly when queue overflow errors  
occur. Queue depth is also adjusted downwardly when  
other initiators are first detected on the fibre  
20 channel network, so as to allow those initiators to  
access the node. The queue depth is adjusted upwardly  
when a predetermined period elapses without queue  
overflow errors.

Nodes that have successfully accessed a target  
25 node without queue overflow errors automatically try  
successively higher values of queue depth until they  
find a depth for optimum performance. Further, queue  
depth is automatically reallocated to active nodes as  
other nodes become idle.

30 In order to limit dispatch of frames over  
intermittent or problem paths, with attendant errors  
and retries, freshly repaired and problem links and

their associated paths are given probationary status. Frames are preferentially dispatched over non-probationary paths if any exist, if no non-probationary paths exist a probationary path may, in  
5 desperation, be returned to normal status. If no probationary paths exist, failed links are tested to determine if they have been repaired; a repaired path may then be activated bypassing the probationary state.

10 Probationary paths that function without error for a preset period of time are returned to normal status, where they may once again be subjected to normal use. Probationary paths that record excessive error counts are returned to failed status, unless  
15 these errors occur within a few seconds of another device logging into a fibre channel fabric; this prevents marking paths failed solely because of transient errors induced by nodes bringing their laser diodes online.

20 The foregoing and other features, utilities and advantages of the invention will be apparent from the following more particular description of a preferred embodiment of the invention as illustrated in the accompanying drawings.

#### 25 BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is an illustration of a Switched Fibre Channel Fabric having several nodes with multiple paths between, and showing drivers and topology databases;

Figure 2, a block diagram of a machine that may run a driver and topology database of the present invention and serve as a network node;

5 Figure 3, a block diagram illustrating where a local network information database resides in a system running on a network node;

Figure 4, an illustration of a data structure for a local network information database as maintained on a node of the present invention;

10 Figure 5, an illustration of a fibre channel frame, or packet, header, as is known in the art;

Figure 6, a diagram of a node record of the topology database;

15 Figure 7, a diagram of a device record of the topology database;

Figure 8, a diagram of a path link of the topology database;

Figure 9, a diagram of a node link of the topology database;

20 Figure 9A, a diagram of an HBA-port record of the topology database;

Figure 10, a flowchart of dynamic queue depth throttling operations performed upon completion of an exchange;

25 Figure 11, a flowchart of dynamic queue depth throttling performed periodically;

006433E 000100

Figure 12, a flowchart of dynamic queue depth throttling operations performed when another initiator node logs into the storage area network;

5 Figure 13, a flowchart illustrating some steps of probationary repair of unstable or failed links

Figure 14, a flowchart illustrating actions taken when a path is detected repaired; and

Figure 15, a flowchart illustrating actions taken when the first frame of an exchange is transmitted.

10

## DETAILED DESCRIPTION

### Context In Which The Invention Operates

15 A Fibre-Channel network (Figure 1) has a first node 100 that comprises a computer 102, a software or firmware driver 104 operating in either the computer or in a dedicated processor of a fiber channel interface card, and a fibre channel port 106. The network also has a fibre channel fabric 108 comprising one or more switches 110 and 112, a plurality of links 114, 116, 118, 120, 122, and 124, and at least one  
20 storage node 126 and 128. The network may also have one or more additional computer nodes, such as node 130.

25 The driver 104 maintains a network topology database 132. This database 132 contains information regarding the topology of the network as is useful for routing packets and determining destinations for valid packets.

Typically, driver 104 and topology database 132 reside in memory of the computer system of a node.

For example, portions of the driver 104 and topology database 132 may reside in main memory devices 200 (figure 2), cache 202 and 204, or disk memory 206 of the computer system. The driver 104 may execute in a first 208 or an additional 210 processor of the node. Typically, one or more fibre channel N\_ports (node ports) are located on a host bus adapter (HBA) 212 attached to the computer system. Portions of the driver 104 and topology database 132 may also reside in local memory (not shown) of the HBA 212, for execution by an embedded processor (not shown) of the HBA. There may be an additional HBA 214, or an additional port of the HBA 212. The one or more levels of cache memory 204, main memory 200, local disk 206 of the node, and local memory of the HBA, together form a memory system of the node, which stores the driver and the network topology database.

Whenever a process 300 (Figure 3) executing on the first 208 (Figure 2) or additional 210 processor of the node generates an I/O request 302, the request is processed by the operating system 303 to determine which driver should handle the transaction. The request may also be translated into block I/O requests 304 for accessing specific blocks of storage, including blocks of directories as well as blocks of data files. Those block I/O requests accessing fibre-channel-accessible devices are passed to the driver 104, which breaks these transactions into packets 305 for the indicated transaction. Packets 305 are processed by the driver 104 to include a valid header, and dispatched to one of one or more ports 306 for dispatch into the network fabric 308 or arbitrated loop.



## Network Topology Database Structure

### HBA Port List

Dispatch of packets to the one of one or more ports is controlled by information from the network topology database 132. Within this database as detailed in Figure 4, is maintained an HBA port list 400. Within this list is a pair of head and tail pointers 402 and an HBA Port Record (404 and 406) for each N\_Port of an HBA on the system.

Each HBA Port Record (404 and 406) incorporates port status information. The HBA Port Record also incorporates the timers and failure counters necessary for implementing probationary paths.

Additional HBA port records may be added as ports are discovered or as hot-pluggable ports are added to the system.

### Node Links

Each HBA port record, such as HBA port record 404, contains head and tail pointers to a linked list of Node Links 408 and 410. Each Node Link 408 contains a pointer to a node record, such as node records 412 and 414. The node record contains routing information required for dispatch of packets over the network from an associated hardware HBA port 212 or 214 to a specific target port of a target node of the storage area network. Target nodes of the storage area network may be storage nodes, such as storage nodes 126 and 128 (Figure 1), or may be processor nodes such as node 130. Each target node has one or

more NL\_ports, L\_ports, or N\_ports 136, 138 and 140, through which it attaches to the fabric 108.

### Node Records

Each node record 412 and 414 has information relevant to a target node of the network. Each node record reachable through pointers in one or more node links, node record 414 is accessible through node link 410 and 416. The node records, including node records 412 and 414, are linked together as a linked target node list 417 accessible through target node list head and tail pointers 418.

### Device Links

Each node record also contains pointers to a list of device links, such as device links 420, 422, 424, and 426, associated with that node record. As illustrated, node record 412 has pointers to a device link list containing device links 420 and 422, while node record 414 has pointers to a device link list containing device links 424 and 426. Each device link points to a device record, such as device records 430, 432, and 434.

### Device Records

Device records, such as device records 430, 432, and 434, include information relevant to individual devices. Each device may be accessible through more than one node or port. For example, a RAID disk array 142 (figure 1), that may be accessed through two RAID controllers having ports 138 and 140, would have a corresponding device record accessible through multiple node records and device links.

The device records 430, 432, and 434 are linked together with a set of head and tail pointers 436 into a searchable device list 438.

### **Path Links**

5        Each device record, such as device records 430, 432, and 434, also has pointers to associated lists of path links 440, 442, 444, 446, and 448. Each path link points to a node link, such as node links 408, 410, or 416, of the topology database; and corresponds  
10      to a physical path through the network from an HBA port to the target node and device. Each path link also has status information about the corresponding path.

15      The network topology database of Figure 4 as drawn does not reflect the specific structure of the example network of Figure 1. Each figure is by way of example, showing the types of records and their linkages, or nodes and devices; the topology database is configured automatically to represent any of a  
20      large variety of possible networks.

### **Detail of Network Topology Database Structure**

#### **Node Record Detail**

25      A node record, such a node record 412, is shown in more detail in Figure 6. Referring to Figures 6 and 4, included in the record are forward and reverse list pointers 600 for linking the record to the target node list pointers 418 and to other node records such as node record 414. There are also device-link list pointers 602 for accessing the list of device links,  
30      such as device links 420 and 422. The device-link

list pointers 602 are therefore pointers through which associated device records may be located.

Each node record also has node information 604 about the corresponding node of the network, including node identity information. The node identity information includes the destination ID (D\_ID) 606 required for routing frames to the node, and node identifying information including a globally unique identifier 608 for the node and port. This identifier is unique among all the nodes of the storage area network. The node identity information may incorporate a node name 610. The globally unique identifier 608 may comprise the node name 610 and a port name associated with the HBA port record through which the node record is reachable since the combination of port name and node name is unique in the network. The node identity information permits determination of whether nodes logging in through a port have already been seen through another port. There may also be information for use in higher levels of protocol, such as a SCSI address 612 for use with SCSI-over-Fibre-Channel protocols.

The node information also has a number of fields used to manage command queues. These include a maximum queue depth 620 for the node, a current queue depth 622, a number of pending operations 624, a timer 626, and a queue depth decrement field 628 for logging the time of the most recent queue depth reduction due to a queue depth refusal.

### Device Record Detail

5 A device record such as device record 430 has forward and reverse pointers 700 (Figure 7) through which it is linked to the device list pointers 436 and other device records. There are also path link list pointers 702 to any associated list of path links, and device information 703. Included in the device information 703 is device identification 704, including a device name for user access to the device.

10 The device identification 704 may also include unique device identification, such as device serial numbers. The device information 703 also includes any logical unit number 706 needed to reference the device, device type, status 707, and other device specific

15 information 708 of interest to the driver.

### Path Link Detail

Each path link, such as path link 440, has list pointers 800 (Figure 8) for linking it with the associated path link list pointers 702 in the

20 associated device record and to any other path links associated with the device. The path link also has path information 804, including path status 806, and redundancy mode information 814. There is also a pointer 816 to an associated node link and a reverse

25 pointer 818 pointing back to the associated device.

Path status 806 includes information indicating which path, of possibly multiple paths, is an active path on which frames should be sent, and which paths are standby paths. It may also include path failure

30 flags.

## 5

10

15

## 20

25

## Network Topology Database Operation

### Topology Maintenance

5 Fibre channel networks are dynamic; nodes may appear or disappear as the network operates, paths may fail or be repaired, and devices may appear or disappear. Host bus adapters may be enabled, added, disabled, removed, or replaced while the system is running, as may individual links. Consider the case when a target node logs into the fabric, and appears to an HBA port 406 (Figure 4). In this event, the target node list 417 is searched from head and tail pointers 418 for a node record having the same node identity. If a node record, such as record 414, is found, node link 416 is added to the database to indicate that the node can be reached through the relevant port 406 as well as through any other port it is reachable through, such as port 404. If no matching node record is found, a new node record is created and linked with a new node link. An inquiry is made of the new node to determine what devices it supports - if any.

25 Similarly, when a node reports the devices it supports, the device list 438 can be searched for a device record corresponding to each device for the same device appearing through any other node. If a matching device record is found, new device links are created to link the node record to the device record and a new path link is created to indicate the new path through the node; if no matching device is found new device links are created and linked to a new device record that is linked into the device list.

```

5      The device and node lists are empty of device
      structures when the driver is started, device
      structures are added to the device list as the driver
      logs into the network and discovers its topology.
      Each device record also has forward and reverse list
      pointers to facilitate dynamic addition and deletion
10     of device structures to the list.

```

## Frame Dispatch With The Topology Database

15           Fibre channel frames, or packets, are known to  
have a structure as illustrated in Figure 5. Each  
frame has a header 500 that includes a destination  
identification field (D\_ID) 502, a source  
identification field (S\_ID) 504, a routing control  
20   field (R\_CTL) 505, and a frame type 506. There are  
also sequence count (SEQ\_CNT) 508, indicating where  
the frame is in a sequence, originator exchange  
identity (OX\_ID) 510, sequence in transaction identity  
(Seq\_ID) 512 and receiver exchange identity (RX\_ID)  
25   514 fields for identifying a frame of an exchange. In  
most fibre channel networks, the D\_ID field 502, type  
506, and routing control fields 505 influence the path  
followed by the packet from source to destination.  
The S\_ID field 504 influences the path followed by any  
30   reply frame.



Consider the case where a process desires access to a specific device by name rather than with a tag. The device may be located by searching from the pointers 436 (Figure 4) of the device list 438 of the network topology database for a device record having the desired device name, such as device record 432. A pointer to this record may be returned as a handle or tag to the calling routine to expedite later access. The list of path links 442 and 444 associated with this device record 432 is then searched for a path link having active status; the associated node link 416 may then be located. Alternatively, an active path pointer 710, pointing to a currently active path link, may be stored in the device record. If no path link has active status, a path link having standby or probationary status may be made active and used instead.

Following the node link pointer 816 of the path link 416, through the node record pointer 904 of the node link, the associated node record 414 is found. From the node record is read the outbound D\_ID, and other relevant information, for packets intended for the specific device.

From the node link 416, pointer 902 is followed to the associated HBA port record 406 to determine the port on which the packets intended for the device should be transmitted, and thereby identify the port that should appear in the S\_ID field of the frame.

Once the D\_ID and S\_ID fields of the frame are determined, along with any relevant routing control information, the frame is dispatched into the fabric or onto the arbitrated loop.

### Dynamic Control of Command Queue Depth

Target nodes often have a command queue associated with them. These nodes may receive and buffer a finite number of queueable commands, which are then executed. As known in the art, these commands need not be executed in order. When an attempt is made to queue more commands than the command queue can hold, excess command exchanges are terminated with queue overflow errors - these commands generally must be re-queued by the initiator node.

Command queue overflow errors are limited through a dynamically-adjusted command queue depth 622 (Figure 6) and a number of pending operations 624 maintained in each node record. The maximum queue depth 620 is initialized to a maximum queue depth suitable for the type of the target node when the node logs into the target node and the node record is created; a minimum queue depth (not shown) is also established for each node. The current queue depth 622 is set to the maximum of the minimum queue depth or the maximum queue depth 620 divided by the number of initiators visible to the HBA port. The number of initiators 962 seen logged into the network and visible to the HBA port then maintained in the HBA port record.

Each time a new I/O operation is requested, the number of pending operations 624 is compared to the current queue depth 622 for the target node of the operation. If the new operation is of a queueable type and will cause the current queue depth 622 to be exceeded, the new operation is linked to a delayed operation queue. These operations will be performed when currently executing operations complete or abort.

If the new operation will not cause the current queue depth 622 to be exceeded, the number of pending operations 624 is incremented, and frames are transmitted as required to initiate the operation.

5           When I/O operation exchanges terminate 1000  
(Figure 10), frames returned by the target node are  
tested 1002 for queue overflow errors. If a queue  
overflow error occurs, its time is compared to the  
last queue decrement time 628, the prior contents of  
10 the last queue decrement time being used to permit  
downward adjustment of queue depth at a maximum rate  
of once every five seconds. If downward adjustment of  
queue depth is done, the last queue decrement time 628  
is updated. The current queue depth 622 is checked  
15 1004 to determine if it is already at the minimum  
value. If a queue overflow error occurs with the  
queue depth at other than the minimum value, the  
current queue depth 622 is adjusted downwardly. The  
downward adjustment is either by five 1010, or to the  
20 minimum value 1006 or 1008, depending on the value of  
the queue depth and the prior contents of the last  
queue refusal variable.

A timer triggers periodic reexamination of  
current queue depth. Upon expiration of the timer,  
25 the last queue decrement time 628 is checked 1100  
(Figure 11) to determine if the current queue depth  
was decreased since the timer last expired. If not,  
and the current queue depth 622 is less than maximum  
1110, the current queue depth 622 is adjusted upwardly  
30 by five 1120 or to maximum 1125, and the timer is  
restarted.

In an embodiment, the timer expiration interval is also dynamically adjusted between a minimum and maximum value during operation. The timer interval is decreased 1135 when it is greater than minimum and current queue depth 622 is adjusted upwardly; the timer interval is increased 1140 when it is less than maximum and no adjustment to the queue depth 622 has been made since the most recent previous timer expiration. Timer expiration interval decreases serve to allow for rapid convergence upon optimal performance, timer expiration interval increases serve to avoid excess processing time when the queue depth is likely already at the maximum usable queue depth.

In order to provide queue capacity to compute or initiator nodes that have just logged into the storage area network, a list (not shown) of known initiators is maintained in the network topology database 132. The node on which the network topology database 132 resides can see some logins by other initiators to the storage area network.

Each time a login by a new initiator node to the fabric is seen, where the new initiator node has access to a target node, the target node list 418 of the network topology database 132 is searched 1200 (Figure 12) to find any node record, such as node record 412, that is likely to be addressed by that newly logged-in initiator node.

If a node record is found that may be a target node of the newly logged-in initiator node, the current queue depth 622 of the node record is reduced 1202 to the greater of the minimum queue depth for that node and the maximum queue depth divided by the

number of initiators 962 now known to be logged into the associated HBA port.

Since the current queue depth 622 is maintained in the node record, there are one or more separate  
5 current queue depth variables for each target node of the network. These variables are adjusted independently; different target nodes may therefore have different current queue depth 622 associated with them. Similarly, there are separate numbers (or  
10 counts) of pending operations 624 maintained in the node records, each corresponding to different pairings of N\_port, L\_port, or NL\_port to a target node of the network.

#### **Recovery and Probation of Failed Paths**

15 Each physical path set up through the network from an initiator node to a target node is associated with an HBA port record and a path link, as heretofore described. The path link has a status flag 806 (Figure 8), which may have the conditions Active, or  
20 Normal, among other possible status conditions. The HBA port record has further status information regarding the path and particularly relevant to status of the physical link between the physical HBA port corresponding to the HBA port record and the fabric.

25 In order to limit dispatch of frames over intermittent or problem links and paths, with attendant errors and retries, freshly repaired and problem paths are given probationary status. Frames are preferentially dispatched over non-probationary  
30 paths if any exist, if no non-probationary paths exist a probationary path may, in desperation, be returned

to active status when they become available for frames.

Each time an error is detected 1300, a routine (Figure 13) is invoked for logging that error against the link and path used for the exchange. Errors are filtered 1302 to exclude errors not potentially related to a failed or unstable link, and evaluated for error severity. Severe errors 1304 may cause the link to be marked failed 1306 immediately in the HBA port record and the HBA port's laser diodes taken off-line. Less serious errors are counted 1308, with the link marked failed or marked unstable only if high numbers of these errors are encountered.

Errors on non-probationary links are identified 1310, and binned into 10-second intervals. Records are maintained 1312 for four consecutive 10-second intervals. Only if errors are found in the current 10-second interval and three immediately previous 10-second intervals is a link considered unstable 1314. If the link is not unstable, retransmission may be attempted or error reported, and the error routine returns 1316.

If the link is determined unstable 1314, the HBA port record is marked unstable 1318 and a search 1320 is made in the topology database for alternate paths to all destinations reached over that link. Failover is done for those paths for which alternate paths exist. If any path does not have an alternate, the error routine returns 1316 without failing the link 1306 or taking the laser diodes off-line.

5

10

20

30

cleared 1342 and the one-minute probation delay 1332 is restarted 1344.

Each time a path is detected repaired (Figure 14), a firmware or software timer associated with that link is cleared 1400 in the HBA port record. The link is also marked probationary 1402.

When an attempt to transmit an exchange is made (Figure 15), the active path pointer 710 (Figure 7) is followed to see if it points to a valid active path. If that pointer does not point to a currently valid and active path, or it points to a failed path, a search 1500 is made among the path links for an active path. If an active path is found, the associated frames are transmitted on that path 1502, this may be an active path over an unstable link if no alternate path exists. If no path is marked active, a search 1504 is made for one or more paths marked normal. If one or more paths marked normal exist, one of those paths is selected, marked active 1506, and used to transmit 1508 the frame. If no path marked Normal exists, a search 1510 is made for one or more paths associated with a link marked probationary. If one or more paths marked Probationary exist, one of those paths is selected, marked active 1512, and used to transmit 1514 the frame. Any associated one-minute probation timer or five-second error login window timer is cancelled. In this way, a Probationary path may be made active in a desperation recovery.

If no active, normal, or probationary path is found, a search is made 1516 for a path taken offline as unstable or failed. If none is found, the exchange attempt is rejected 1518 since there is no path on



which it can be sent. If such an offline path is found, a repair detect attempt 1520 is made; this requires bringing the laser diodes back online and testing the associated link. If the link tests good, any required logins 1522 are done, the path is marked active 1524 in the path link, any other paths through the link are marked normal, and the path is used 1526 to transmit the frame. In this way, an offline link may be made active in a desperation recovery. If the link tests bad, the laser diodes are taken back offline 1530 and the frame is rejected 1518 since there is no path on which it can be sent.

### Conclusion

While the invention has been particularly shown and described with reference to one or more embodiments thereof, it will be understood by those skilled in the art that various other changes in the form and details may be made without departing from the spirit and scope of the invention.

For example, the order of fields within the node record, path link, and device record may be readily altered for a particular embodiment of the invention. Similarly, the reverse pointers of the node record and the device record are present for convenience in inserting and removing records from the associated lists, while these confer advantages it is possible to construct a functional system with only forward links. If the number of devices addressed through a single node is limited, device links and the associated device link list pointers of the node records may be replaced by a plurality of device record pointers in the node records.

Node records have been described as accessed through node link records, device records through device link records, and paths through path link records. Each link record has pointers for linking into a linked list and pointers to associated node and device records. It is also possible to use a long fixed-length or a variable-length HBA-port record, node record, or device record; the longer or variable length record containing multiple pointers to other node, device, or node records of the database. Using long fixed-length or variable length records may, however, limit the number of nodes, devices and paths trackable by the database. Further, excessive delays and memory fragmentation may be encountered as variable-length record sizes are adjusted while adding nodes and devices to the database.

While the above description has focused on dispatch of frames from a node, switches of a switched fabric also have processors and must maintain a network topology database. Packets received by a switch may be routed on preferred or alternate paths by such a switch according to information in a network topology database of the present invention. Each such switch may also implement the probationary period of the present invention should paths through that switch be repaired. Switches implementing probationary periods on repaired paths may also promote probationary paths to normal operation in desperation when no alternative, normal status, paths remain available.

A computer program product is a computer readable memory device or media having recorded upon it computer readable instructions.

It should be understood that this description has been made by way of example, and that the invention is defined by the scope of the following claims.

004433, 000100

**CLAIMS**

What is claimed is:

1           1. An initiator node for a storage area network,  
2           the node intended to be coupled over a storage area  
3           network to at least one storage node having command  
4           queue capability, the initiator node comprising:

5                 at least one processor capable of executing  
6                 instructions;

7                 a memory system having stored a current queue  
8                 depth, and a maximum queue depth associated with each  
9                 storage node of the at least one storage node;

10                wherein the memory system stores

11                instructions for initializing the maximum queue  
12                depth for the at least one storage node to a value  
13                dependent on a type of the at least one storage node,

14                instructions for limiting the number of commands  
15                queued to a storage node of the at least one storage  
16                node to the current queue depth associated with the  
17                storage node, and

18                instructions for dynamically adjusting the  
19                current queue depth associated with the storage node  
20                based upon queue refusals generated by the storage  
21                node and the maximum queue depth associated with the  
22                storage node.

1           2. The initiator node of Claim 1, wherein the  
2           instructions for dynamically adjusting the current  
3           queue depth include instructions for adjusting the  
4           current queue depth associated with a storage node  
5           downwardly when the current queue depth is greater

6 than a minimum queue depth and the storage node  
7 refuses to queue a command issued by the initiator  
8 node, and for adjusting the current queue depth  
9 upwardly when the current queue depth is less than the  
10 maximum queue depth associated with the storage node  
11 and that storage node has not refused to queue any  
12 commands issued by the initiator node for a determined  
13 period of time.

1 3. The initiator node of Claim 2 wherein the  
2 instructions for dynamically adjusting the current  
3 queue depth include instructions for monitoring  
4 logins, and for adjusting the current queue depth  
5 downwardly when a login by an additional initiator  
6 node is recognized.

1 4. In a storage area network node, a method of  
2 controlling a maximum number of commands queued to a  
3 storage node comprising the steps of:

4 maintaining a maximum queue depth associated with  
5 the storage node and initializing the maximum queue  
6 depth according to a type of the storage node;

7 maintaining a count of outstanding commands that  
8 have been submitted to the storage node;

9 maintaining a current queue depth associated with  
10 the storage node and initializing the current queue  
11 depth to a value not less than a minimum queue depth  
12 nor greater than the maximum queue depth;

13 holding commands for later submission to the  
14 storage node if the count of commands that have been  
15 submitted is greater or equal to the current queue  
16 depth;

17        adjusting the current queue depth associated with  
18        a storage node downwardly when the current queue depth  
19        is greater than the minimum queue depth and the  
20        storage node refuses to queue a command issued by the  
21        initiator node; and

22        adjusting the current queue depth upwardly when  
23        the current queue depth is less than the maximum queue  
24        depth associated with the storage node and the storage  
25        node has not refused to queue any commands issued by  
26        the initiator node for a first predetermined period of  
27        time.

1        5. The method of Claim 4, further comprising the  
2        step of adjusting the current queue depth associated  
3        with the storage node downwardly is permitted to occur  
4        no more than a predetermined number of times in a  
5        second predetermined period of time.

1        6. The method of Claim 5, wherein the  
2        predetermined number of times is one.

1        7. The method of Claim 5, wherein the first  
2        predetermined period of time is adjusted dynamically.

1        8. The method of Claim 5, further comprising the  
2        steps of:

3        maintaining a second maximum queue depth  
4        associated with a second storage node and initializing  
5        the second maximum queue depth according to a type of  
6        the second storage node;

7        maintaining a second count of outstanding  
8        commands, indicating commands that have been submitted  
9        to the second storage node and have not completed; and

10           maintaining a second current queue depth  
11           associated with the second storage node and  
12           initializing the second current queue depth to a value  
13           not less than a minimum queue depth nor greater than  
14           the second maximum queue depth.

1           9. The method of Claim 5, wherein the step of  
2           initializing the current queue depth initializes the  
3           current queue depth to a value dependent upon a number  
4           of initiator nodes known to be logged-in to the  
5           storage area network.

09643575.082400

## 5

10



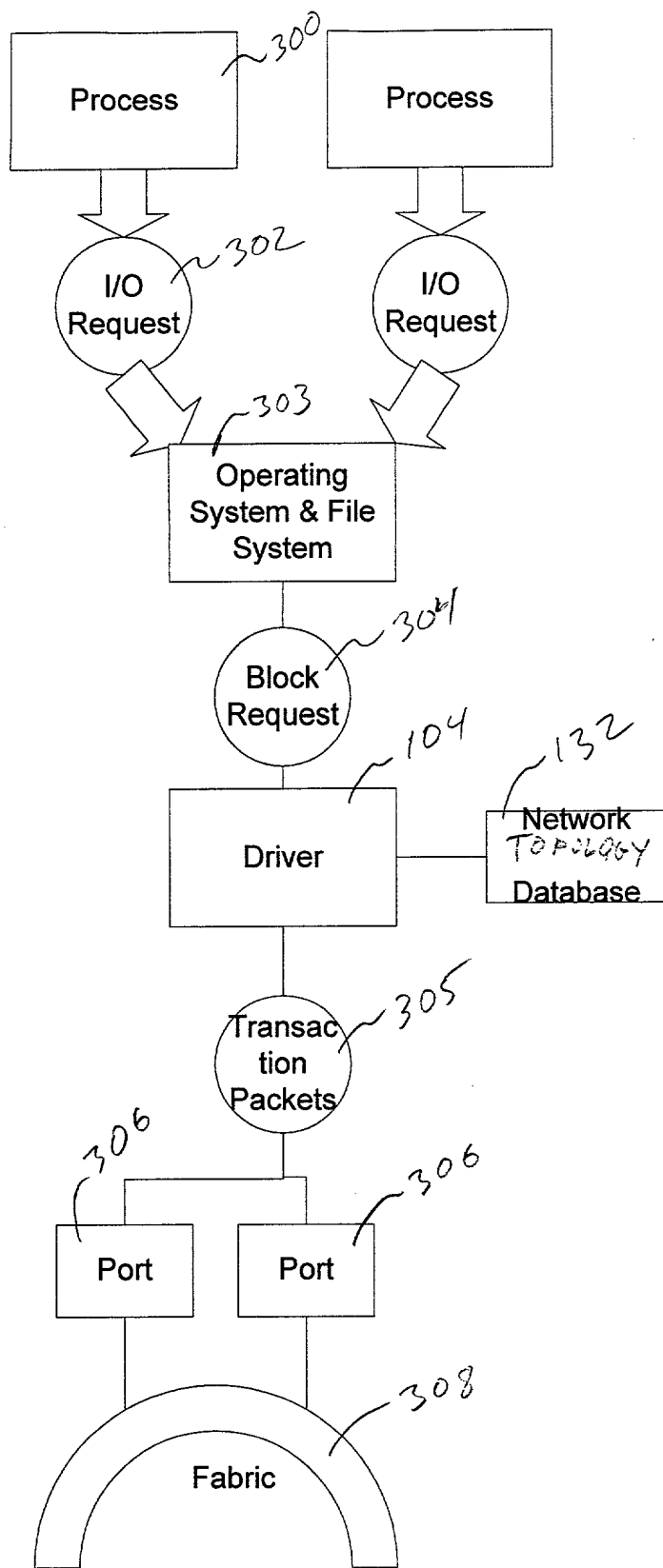
The diagram illustrates a network architecture with two main nodes, Node 1 and Node 2, connected via a central fabric. Node 1 (top left) contains a Computer (102), a Driver (104), a Topology Database (132), and a Port (106). Node 2 (top right) contains a Computer, a Driver, a Topology Database, and two Ports. The central fabric (118) consists of two Switches (114, 122) connected by a FABRIC (118). Below the switches are two Target Storage Nodes: a Disk Server (128) connected to Port 136, and a RAID system (142) connected to two Ports (140, 126). Handwritten annotations include 100, 108, 110, 112, 116, 120, 124, 138, and 140.

### Figure 1

Figure 2 is a block diagram illustrating a computer system architecture. The central component is the **System Bus**. Connected to the System Bus are the following components:

- Optional Additional Processors & Cache** (210)
- 3rd level cache (optional)** (201)
- Second Level Cache** (202)
- Processor** (208)
- Bus Bridge, local bus to PCI** (208)
- Memory Devices** (200) connected via a **Main Memory Controller**
- Video**
- Network Interface**
- Fibre Channel HBA Port** (212)
- Additional Fibre Channel HBA Ports (Optional)** (214)
- Keyboard/Mouse/Serial Port/Printer port**
- DISK MEMORY (OPTIONAL)** (206)

## Figure 2



### Figure 3

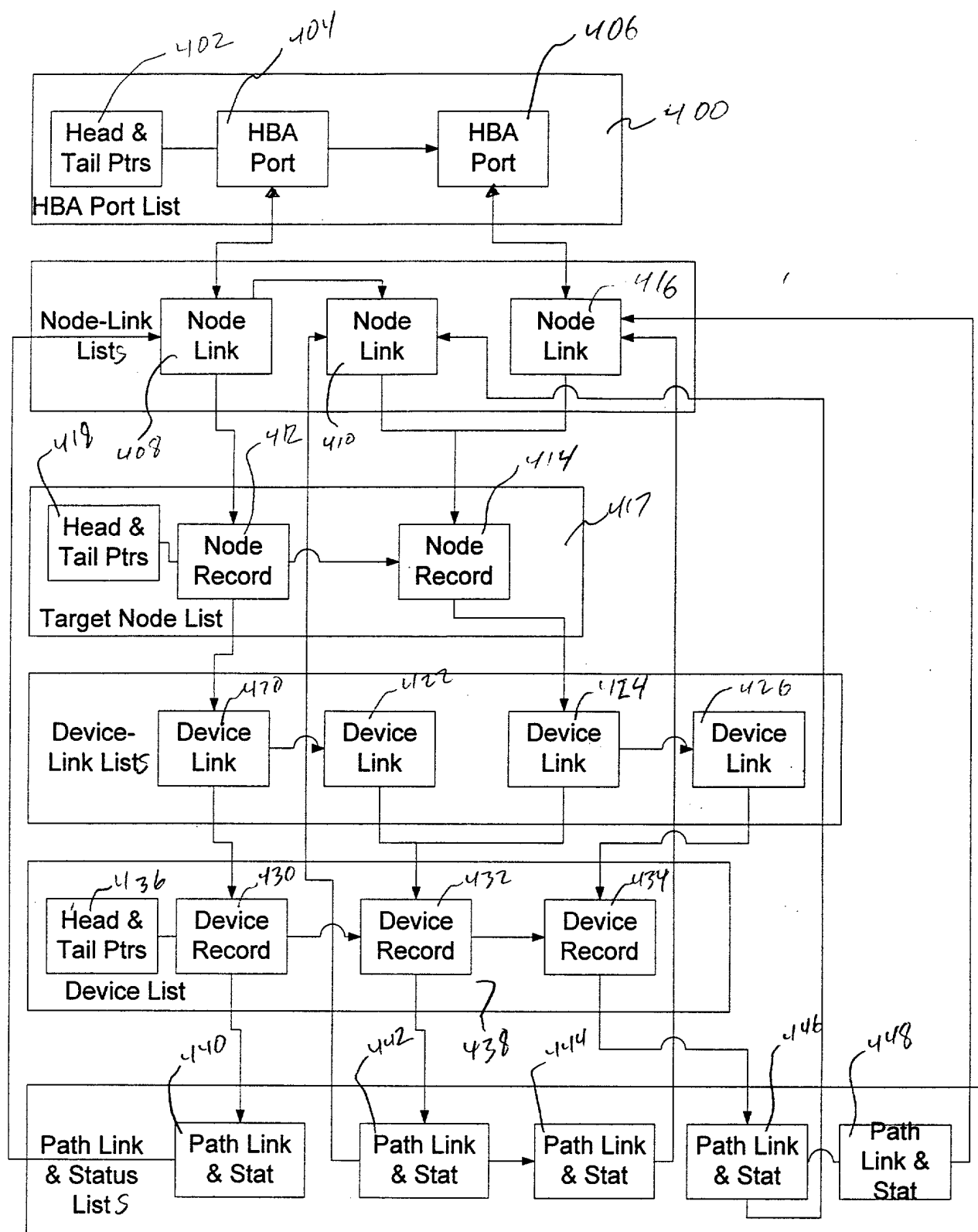
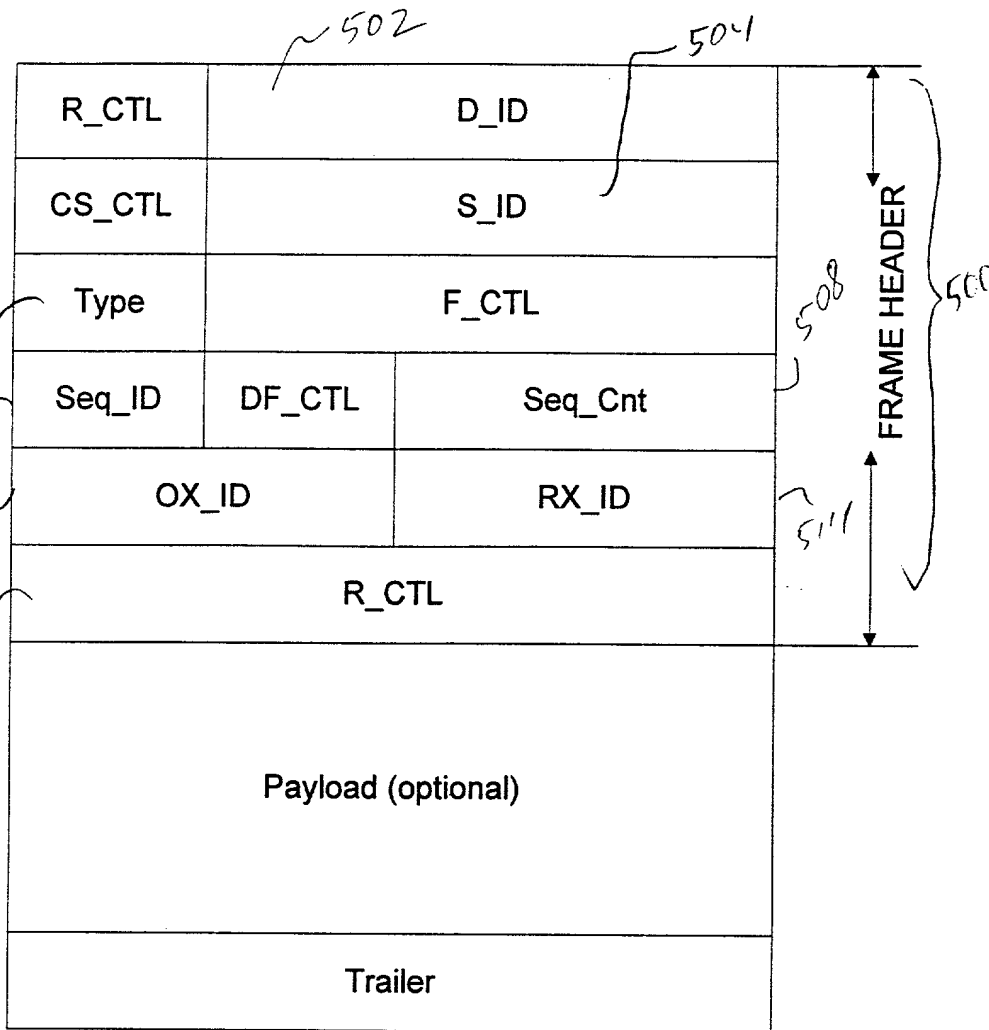


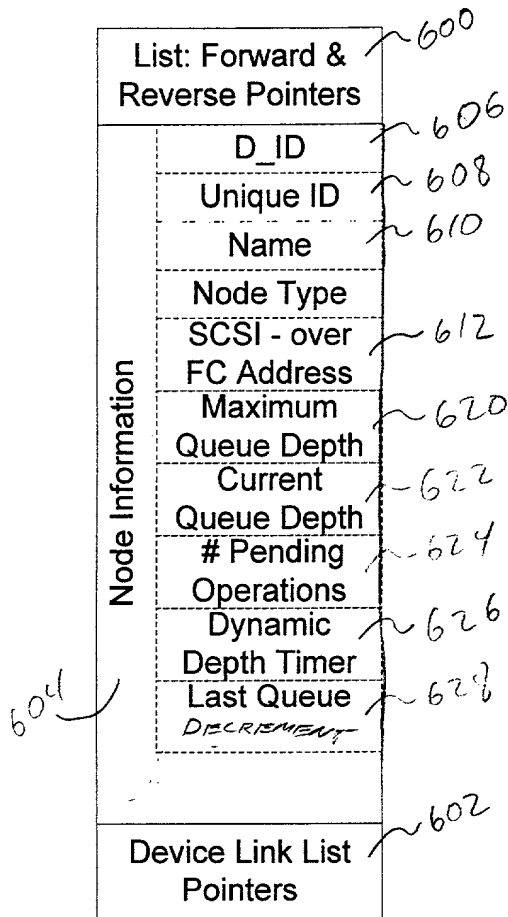
Figure 4



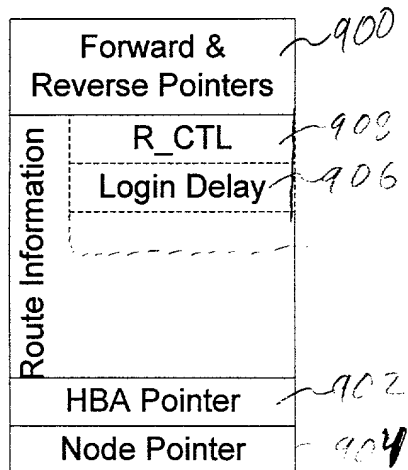
Frame Structure

Figure 5

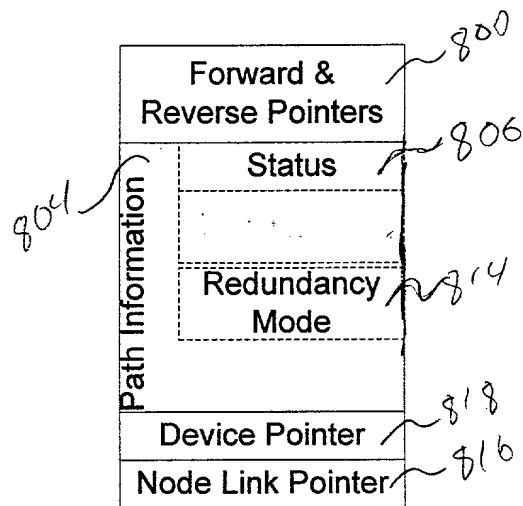
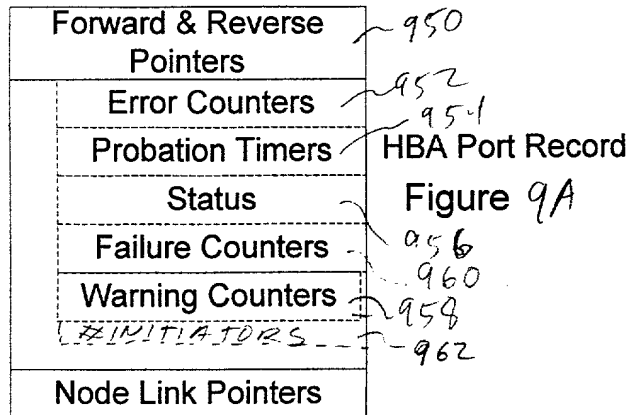
KNOWN ART



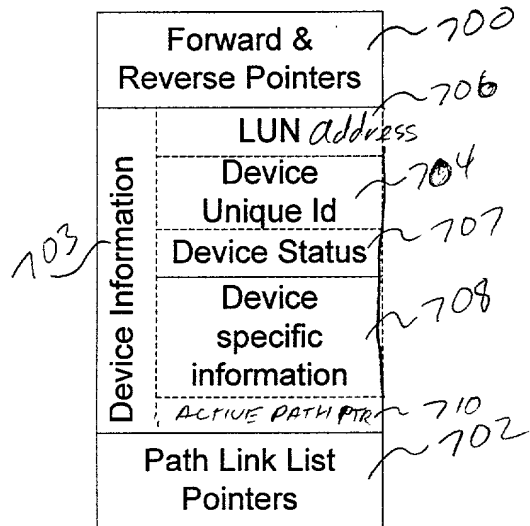
Node Record Figure 6



Node Link Figure 9



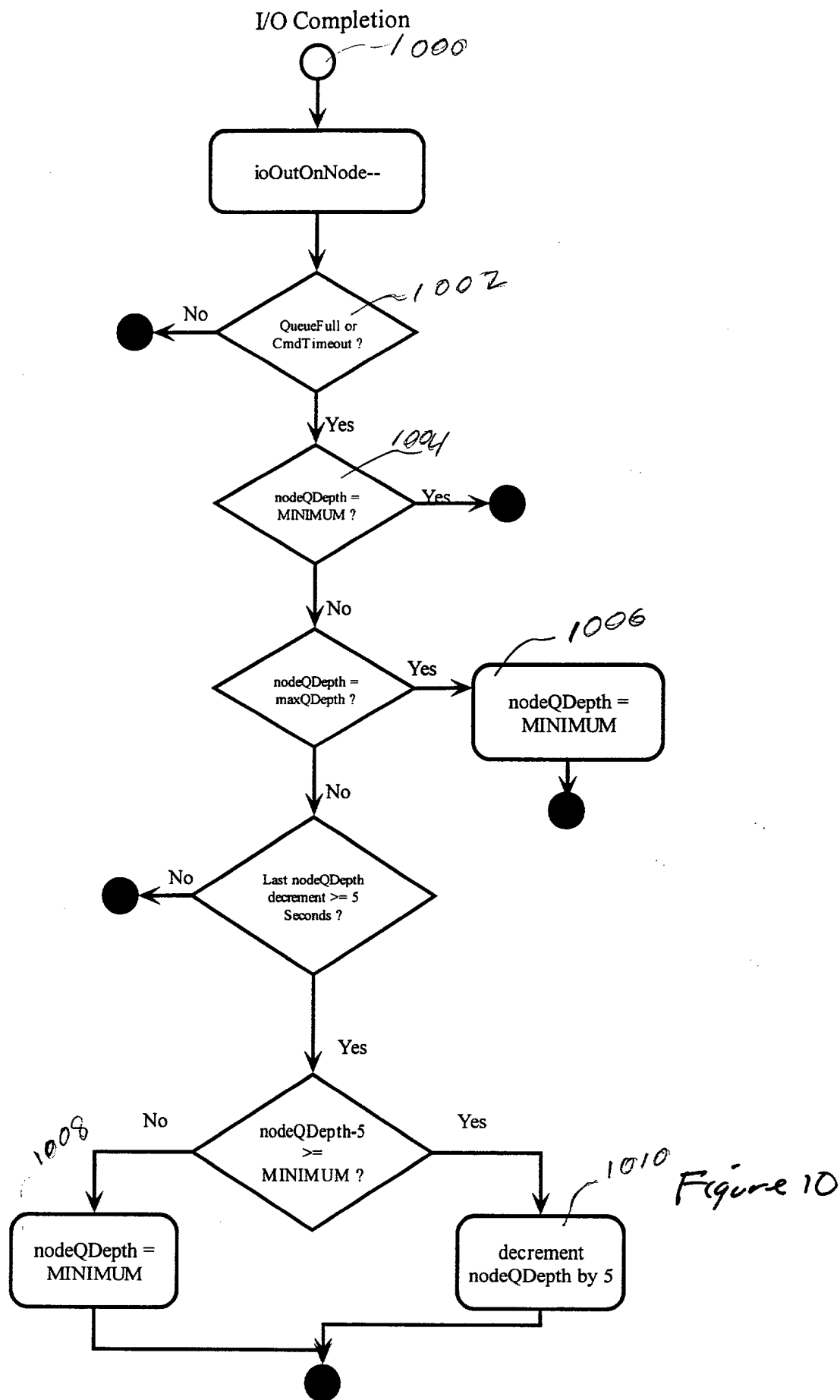
Path Link Figure 8



Device Record

Figure 7

007230" 52524950



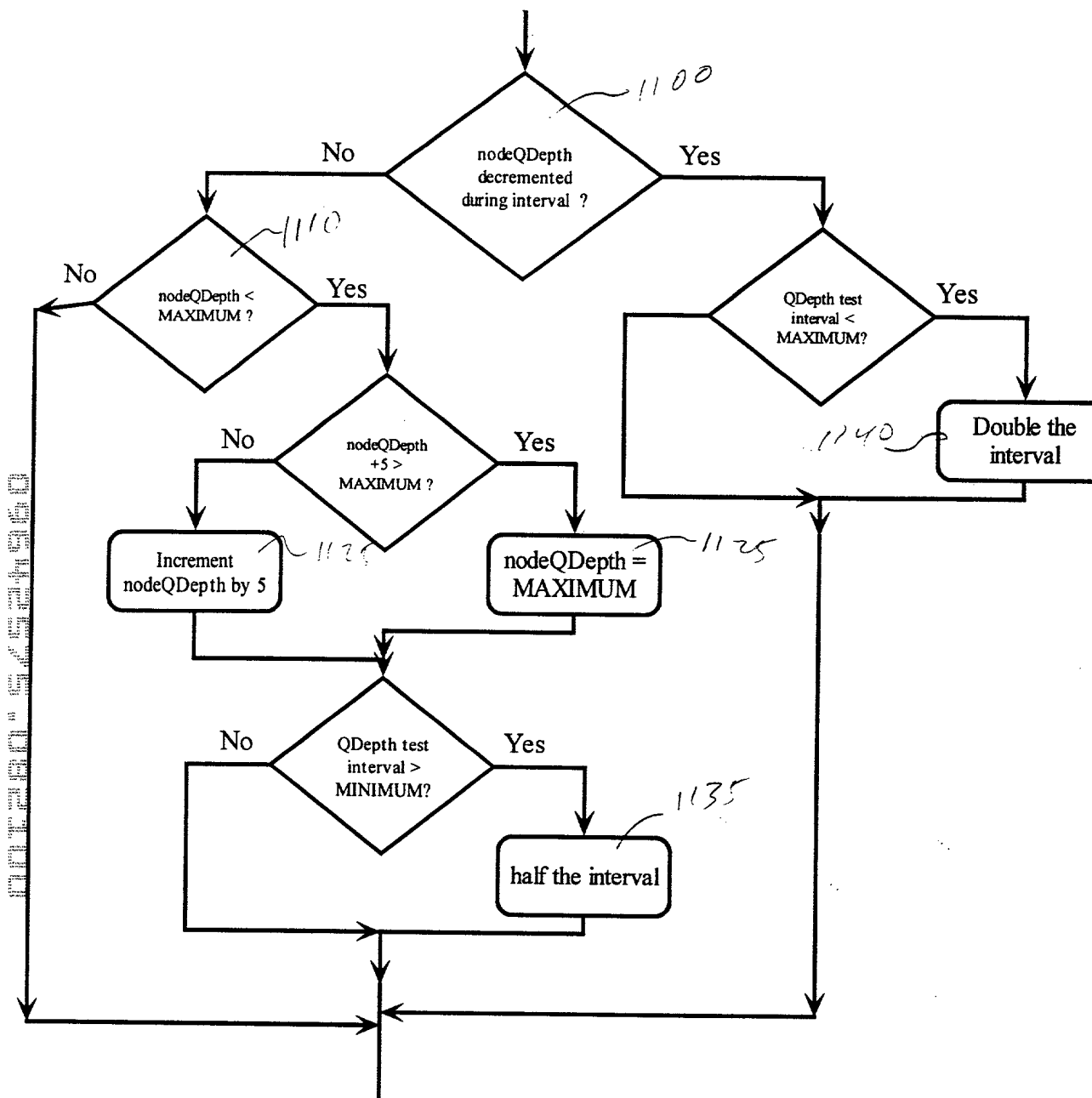


Figure 11



New initiator login

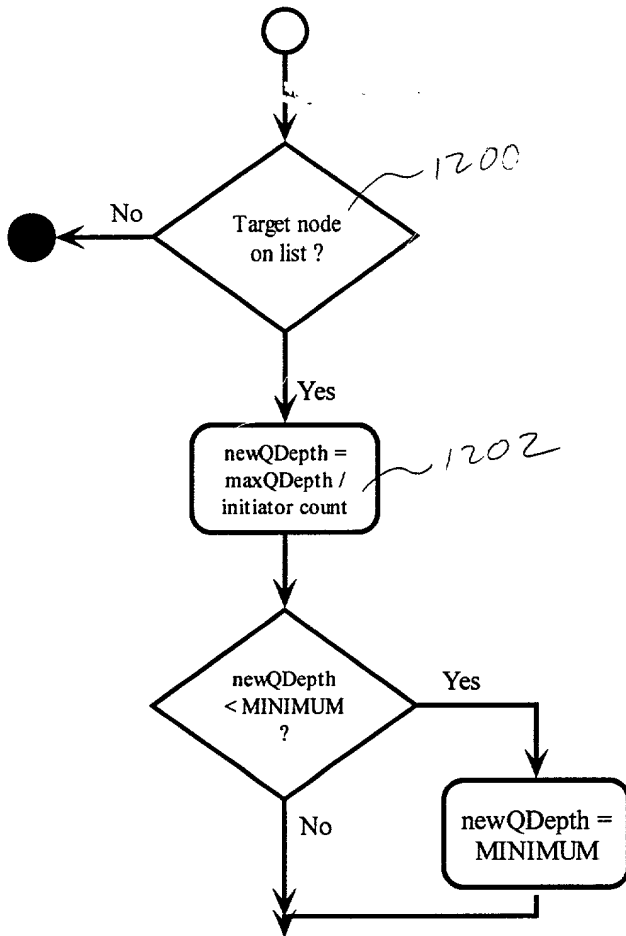


Figure 12

004423-00400

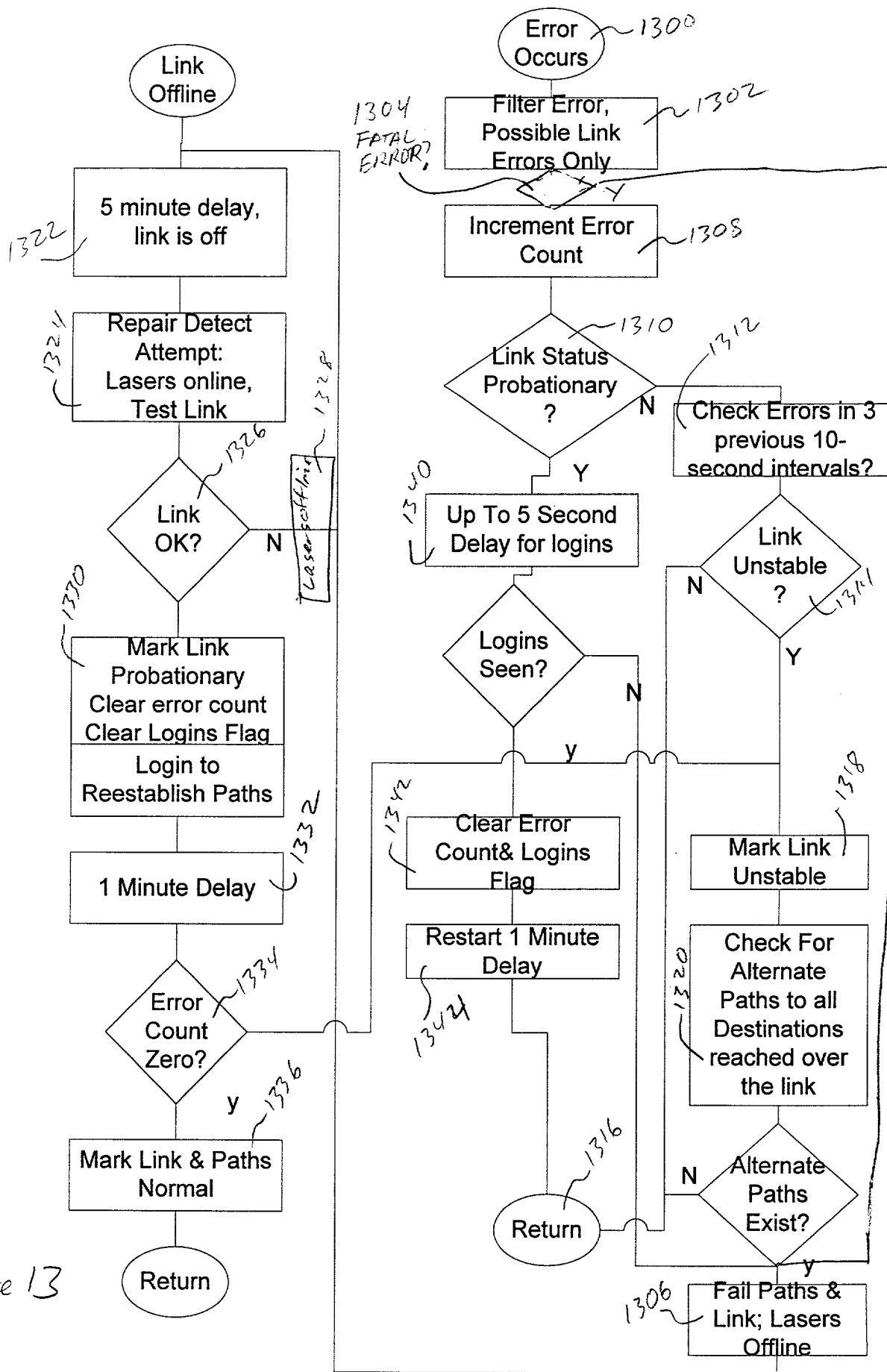


Figure 13

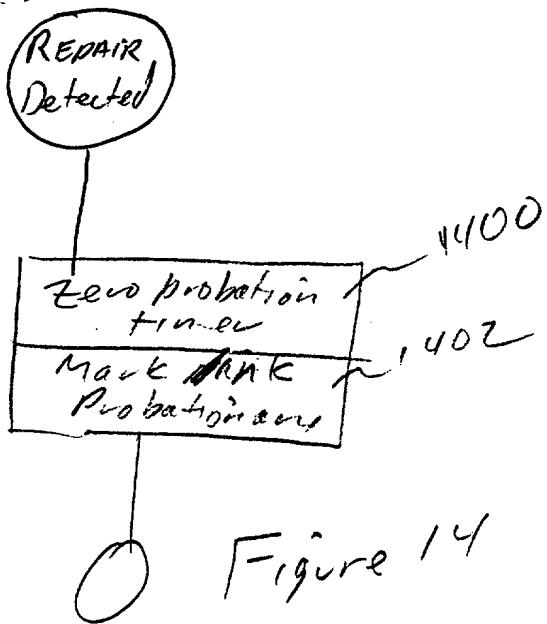


Figure 14

Figure 15

DECLARATION

As a below named inventor, I hereby declare that: my residence, post office address, and citizenship are as stated below next to my name. I believe I am the original, first, and sole inventor (if only one name is listed below) or a joint inventor (if plural inventors are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled:

Apparatus and Method for Dynamic Command Queue Depth Adjustment for Storage Area Network Nodes

as described in the specification ☒ attached or ☐ of patent Application Serial No. \_\_\_\_\_  
filed Herewith and amended on \_\_\_\_\_.

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claims, as amended by any amendment referred to above; that I do not know and do not believe the same was ever known or used in the United States of America before my or our invention thereof, or patented or described in any printed publication in any country before my or our invention thereof or more than one year prior to this application; that the invention has not been patented or made the subject of an inventor's certificate issued before the date of this application in any country foreign to the United States of America on an application filed by me or my legal representative or assigns more than twelve months prior to this application; and that I acknowledge the duty to disclose information of which I am aware which is material to the examination of this application in accordance with Title 37, Code of Federal Regulations § 1.56(a). Such information is material when it is not cumulative to information already of record or being made of record in the application, and

- (1) it establishes, by itself or in combination with other information, a prima facie case of unpatentability of a claim; or
- (2) it refutes, or is inconsistent with, a position the applicant has taken or may take in:
  - (i) opposing an argument of unpatentability relied on by the Office, or
  - (ii) asserting an argument of unpatentability.

I hereby claim foreign priority benefits under Title 35, United States Code § 119 of any foreign application(s) for patent or inventor's certificates listed below and have also identified below any foreign application(s) having a filing date before that of the applications(s) on which priority is claimed:

COUNTRY	APPLICATION NUMBER	Date Filed	Priority Claimed under 35 USC 119
			<input type="checkbox"/> YES <input type="checkbox"/> NO

I hereby claim the benefit under Title 35 United States Code § 120 of any United States application(s) listed below and, insofar as any subject matter of any claim of this application is not disclosed in the prior United States Application, I acknowledge the duty to disclose material information as defined in Title 37, Code of Federal Regulations § 1.56(a) which occurred between the filing date of the prior application and the national PCT international filing date of this application.

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

FULL NAME OF SOLE OR FIRST INVENTOR	INVENTOR'S SIGNATURE	DATE
Richard D. Gunlock	<i>Richard D. Gunlock</i>	8-11-00
RESIDENCE		CITIZENSHIP
12607 Amado Dr., Houston, TX 77065		USA
POST OFFICE ADDRESS		
Same		
FULL NAME OF SECOND JOINT INVENTOR	INVENTOR'S SIGNATURE	DATE
RESIDENCE		CITIZENSHIP
POST OFFICE ADDRESS		
FULL NAME OF THIRD JOINT INVENTOR	INVENTOR'S SIGNATURE	DATE
RESIDENCE		CITIZENSHIP
POST OFFICE ADDRESS		

Diane H. Strong  
NAME: Diane H. Strong  
TITLE: Administrator, Patents  
Authorized To Sign This Document On Behalf Of  
Compaq Computer Corp.